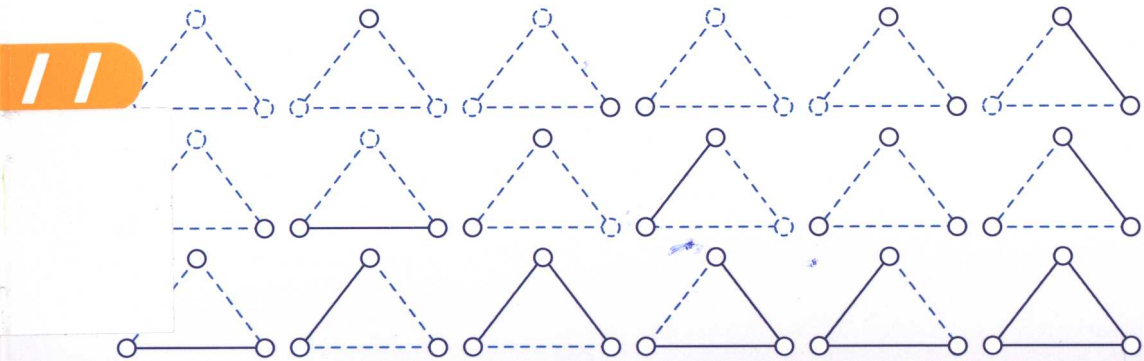
“十二五”国家重点图书出版规划项目

不确定图数据挖掘

Uncertain Graph Data Mining

邹兆年　李建中　著



哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

“十二五”国家重点图书出版规划项目

不确定图数据挖掘

邹兆年　李建中　著

哈尔滨工业大学出版社

内 容 简 介

本书是国际上第一部系统阐述不确定图数据挖掘理论、技术和算法的学术专著。本书系统介绍了不确定图数据挖掘的数据模型、挖掘问题语义以及典型挖掘问题的计算复杂性和算法，具体包括期望语义下的频繁子图模式挖掘算法、概率语义下的频繁子图模式挖掘算法、极大团挖掘算法、紧密顶点子集挖掘算法、可靠子图挖掘算法和聚类算法。

本书适用于从事数据库与数据挖掘理论与技术研究的专业人员及相关高等院校师生阅读。

图书在版编目（CIP）数据

不确定图数据挖掘/邹兆年，李建中著. ——哈尔滨：哈尔滨工业大学出版社，2013.1

ISBN 978-7-5603-3783-8

Ⅰ.①不… Ⅱ.①邹… ②李… Ⅲ.①数据库系统—数据采集—研究 Ⅳ.①TP311.13 ②TP274

中国版本图书馆 CIP 数据核字（2012）第 210636 号

责任编辑 王桂芝 段余男

出版发行 哈尔滨工业大学出版社

社 址 哈尔滨市南岗区复华四道街 10 号 邮编 150006

传 真 0451-86414749

网 址 http://hitpress.hit.edu.cn

印 刷 哈尔滨市石桥印务有限公司

开 本 787mm×960mm 1/16 开 印张 14.75 字数 220 千字

版 次 2013 年 1 月第 1 版 2013 年 1 月第 1 次印刷

书 号 ISBN 978-7-5603-3783-8

定 价 48.00 元

前　言

由于数据获取技术、数据传输技术和数据处理技术的局限性，以及数据隐私保护等原因，不确定性在图数据中普遍存在。如何从大量不确定图数据中自动发现有用的知识业已成为数据挖掘领域的一个研究热点，称作不确定图数据挖掘。近年来，不确定图数据挖掘在模型、理论和算法方面均取得重大进展，在应用方面也发展迅速。尽管如此，目前尚不存在系统介绍不确定图数据挖掘的专业读物。

本书对作者在不确定图数据挖掘领域取得的研究成果进行了全面的梳理和总结，从数据模型、问题语义、计算复杂性和算法几方面系统介绍了不确定图数据挖掘，力图为研究人员提供一个了解不确定图数据挖掘的严谨、易懂的学术读物。

本书共 9 章，可分为两部分。第一部分 (第 1～3 章) 介绍不确定图数据挖掘的基本概念、数据模型、问题语义等理论基础。第二部分 (第 4～9 章) 介绍具体的不确定图数据挖掘问题，证明这些问题的计算复杂性，并提出解决这些问题的算法。第一部分是第二部分的基础，需要首先阅读。在第二部分中，除第 4 章和第 5 章外，其余各章之间相互独立，读者可以根据自身需要按不同顺序阅读。

本书的内容建立在图论、概率论、计算复杂性理论、数据挖掘和算法等理论和技术的基础上，因此需要读者具备上述理论和技术的基础知

识。书中的论述基于严格的数学证明，除一些特殊情况外，所有定理都给出了完整的数学证明。同时，为了使读者更容易理解本书中的概念、定理和算法，书中还提供了大量直观且易懂的例子。

本书由邹兆年和李建中共同撰写，李建中撰写本书第 1 章，邹兆年撰写本书第 2～9 章。

高宏教授阅读了本书初稿，并提出了许多宝贵意见，在此表示衷心感谢。本书的撰写还得到了国家重点基础研究计划 (973 计划) 项目 (批准号 2012CB316200)、国家自然科学基金重点项目 (批准号 61033015) 和国家自然科学基金面上项目 (批准号 61173023) 的部分资助。

由于作者水平有限，疏漏及不妥之处在所难免，望读者批评指正。

邹兆年　李建中

2012 年 8 月于哈尔滨工业大学

目　录

第 1 章　不确定图数据挖掘概述

随着现代化数据采集技术 (如高通量生物实验、无线传感器网络、全球卫星定位系统、社会网络系统等) 的飞速发展，社会各领域中积累了大量用图表示的数据，简称**图数据** (graph data)，如蛋白质交互网络、无线传感器网络拓扑结构、道路网络、社会网络等。这些图数据的规模非常巨大，例如在著名的生物数据库 BioGRID 中，蛋白质交互网络具有 37 万多条边；美国加利福尼亚州道路网络具有 550 万多条边；全球最大的社会网络 Facebook 具有近 6 亿个顶点；万维网具有 130 亿多个顶点。并且，这些图数据的规模还在不断地快速增加。

从规模如此巨大的图数据中人工发现知识显然是不现实的。于是，**图挖掘** (graph mining) 技术应运而生，其目的是从海量图数据中自动发现有用的知识。虽然国内外已经对图挖掘进行了广泛研究，但这些研究几乎全部针对“确定”的图数据，简称**确定图数据** (certain graph data)。然而，现实中的图数据普遍存在**不确定性** (uncertainty)。

本书将系统介绍从含有不确定性的图数据中自动发现知识的理论、技术和算法。这个全新的领域被称作**不确定图数据挖掘** (uncertain graph data mining) 或**不确定图挖掘** (uncertain graph mining)。本章将向读者呈现不确定图数据挖掘的概貌，具体回答以下几个问题：不确定图数据是如何产生的？不确定图数据挖掘与传统的确定图数据挖掘有哪些

本质区别？不确定图数据挖掘面临着哪些新挑战？不确定图数据挖掘领域的主要研究内容是什么？

1.1 不确定图数据的产生

近年来，研究人员发现在图数据的规模不断激增的同时，由于数据获取技术的随机错误与测量误差、数据传输的故障与延迟、多源集成数据的不完整性与不一致性、数据隐私保护等多种原因，大量图数据具有不确定性。

(1) 在生物信息学中，蛋白质交互网络 (protein-protein interaction network，简称 PPI network) 用于记录蛋白质之间的交互作用。蛋白质交互网络被表示为图，其中顶点表示蛋白质，边表示蛋白质交互 [1–5]。由于蛋白质交互的高通量生物检测技术 (如酵母双杂交技术) 存在固有误差，因此实验测得的蛋白质交互是否真实存在是不确定的 [6–9]。著名的生物数据库 STRING [4] 已将这种不确定性信息量化存储于数据库中。图 1.1 给出了一个蛋白质交互网络的片段，其中顶点上的文字是蛋白质的名称，边上的值表示蛋白质交互真实存在的可能性 (数据取自 STRING 数据库)。

(2) 无线传感器网络 (wireless sensor network, 简称 WSN) 是一种由大量分布在地理区域内的传感器节点以自组织方式构成的计算网络，其目的是协作地采集、传输和处理网络所覆盖的地理区域中感知对象的信息，并发布给观察者 [10, 11]。无线传感器网络在国防军事、环境监测、灾害预防、交通管理、医疗卫生、制造业等领域具有广泛应用。无线传感器网络的拓扑结构是一个图，其中顶点表示传感器节点，边表示传感器节点之间的无线通信链路。由于物理世界的干扰、无线传感器网络的动态性、传感器节点的移动性、易失效性和睡眠机制等原因，两个传感

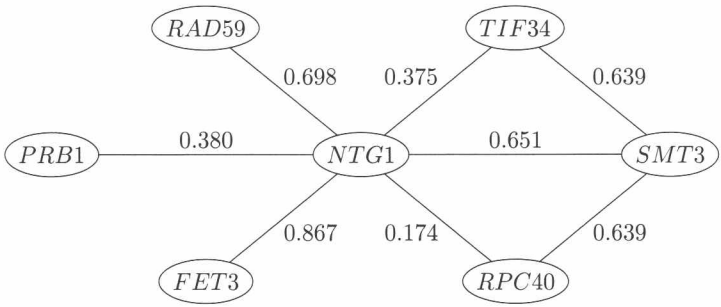


图 1.1　蛋白质交互网络 (数据来自 STRING 数据库)

器节点之间是否真实存在无线通信链路是不确定的 [12, 13]。

(3) 社会网络 (social network) 是一种用于刻画人与人之间社会关系的模型，最初被社会学家用于研究社会团体的组成结构和演化机理 [14]。社会网络可以表示为一个图，其中顶点代表人，边代表人与人之间的社会关系。由于人与人之间的社会关系具有动态性和时效性，因此在社会网络中两个人之间此时此刻是否存在关系是不确定的 [15–18]。另外，出于对个人隐私的保护，社会网络中某些敏感的人际关系会被刻意隐藏起来，或者在社会网络中添加一些无关紧要的社会关系，以混淆敏感的社会关系 [19–23]，这也使得社会网络具有不确定性。

(4) 在智能交通系统中，道路网络 (road network) 被表示为加权图，其中顶点表示道路交叉口，边表示道路，边上的权值表示道路上的交通流量 [24]。由于全球定位系统 (GPS) 的数据传输故障和延迟，当前道路网络中的交通流量数据无法确切反映实际情况，因而具有不确定性。

本书沿用作者在文献 [18, 25–44] 中使用的术语，将这种具有不确定性的图数据称作**不确定图数据** (uncertain graph data)。尽管一些文献 (如 [12, 45, 46]) 将具有不确定性的图数据称作概率图数据 (probabilistic graph data)，但是这种提法是不完全准确的，因为概率只是不确定性的表示方法之一 [47]；对于不能用概率表示的不确定性 [48]，我们不能将这

种图数据称作概率图数据。

1.2 不确定图数据挖掘的概念

各领域中大量存在的不确定图数据中蕴含了大量有用的知识。不确定图数据的巨大规模使得人工分析这些数据并从中归纳和发现知识是不可能的，因此从不确定图数据中自动发现这些有用的知识变得愈发重要和迫切。

(1) 带有不确定性的蛋白质交互网络中不仅蕴含着大量有关蛋白质复合体 (protein complex) 组成的知识，而且还蕴含着这些组成形式存在的可能性 [7]。生物学家经常需要了解蛋白质交互网络中到底蕴含着哪些存在可能性较高的蛋白质复合体，用于指导生物检测实验，以提高实验成功率，节约实验成本。

(2) 带有不确定性的无线传感器网络拓扑结构图中蕴含着传感器节点之间的通信模式，该模式对于设计可靠、实时、自适应的无线传感器网络路由协议具有重要作用 [12]。无线传感器网络管理员经常需要了解可能性较高的频繁通信的路径或子网络，以采取措施，避免因这些路径或子网络上的传感器节点电源耗尽而导致整个网络通信断开。

(3) 人们可以从不确定的道路网络中发现可信度高的交通流量模型，该模型可以在道路网络不确定、数据无法及时更新的情况下，给出可信度高的交通流量估计。这对建立智能交通系统，实现对交通流量的有效控制和疏导具有重要作用。

(4) 社会网络中蕴含着大量社会学、医学所关注的知识，如人与人之间社会关系的形成原理、信息在人与人之间传播的规律、传染性疾病在人群中传播的机理等。

现实应用亟需从海量不确定图数据中高效、自动地发现有用的知

识。这种从不确定图数据中自动发现有用知识的过程称作**不确定图数据挖掘** (uncertain graph data mining)。

不确定图数据挖掘与传统的确定图数据挖掘存在显著区别：

(1) 在数据模型方面，确定图数据挖掘使用图论中的图模型来表示图数据 [49, 50]，然而不确定图数据挖掘不仅要表示图的结构，还要表示图的不确定性，这是图论中的图模型做不到的。

(2) 在问题语义方面，确定图数据挖掘只考虑图的结构信息，而不确定图数据挖掘将图数据的不确定性作为和图的结构同等重要的信息纳入知识的重要性度量，从而提供了更丰富的语义信息，有助于提高挖掘结果的质量。

(3) 在计算复杂性方面，不确定图数据上某一挖掘问题的计算复杂性通常要比确定图数据上同一挖掘问题的计算复杂性要高。

(4) 在知识表示方面，确定图数据挖掘只给出挖掘到的知识的结构及重要性，而不确定图数据挖掘不仅要给出知识的结构和重要性，还要给出不确定图数据中真正蕴含该知识的可能性。

1.3　不确定图数据挖掘面临的挑战

不确定图数据挖掘能够引入更丰富的信息，表达更丰富的语义，发现更可靠的知识，但同时也提出了新的挑战性问题：

(1) 如何表示图数据的不确定性。不确定性是一个广义范畴，其中包含随机性、模糊性、不精确性、不一致性、不完整性和不实时性。不确定图数据挖掘首先需要建立简洁、易用、灵活的不确定图数据模型来表示不同类型的不确定性，实现模型表达能力与易用性之间的平衡。

(2) 如何衡量不确定图数据中知识的重要性。相对于确定图数据，不确定图数据中知识的重要性度量发生了显著变化，需要将图数据的不

确定性纳入考察范围，从而确保从不确定图数据中挖掘出重要且可靠的知识。

(3) 如何设计时间复杂性最小化的不确定图数据挖掘算法。不确定图数据对挖掘算法的时间复杂性提出了更高的要求，需要设计具有更低时间复杂性的挖掘算法。

1.4 不确定图数据挖掘的研究内容

针对不确定图数据挖掘所面临的挑战，国内外纷纷开展了不确定图数据挖掘的研究工作。不确定图数据挖掘现已成为数据挖掘领域新兴的研究热点，目前的研究工作主要涉及以下几个方面：

(1) 不确定图数据模型；

(2) 不确定图数据挖掘问题的语义；

(3) 不确定图数据挖掘问题的计算复杂性；

(4) 不确定图数据挖掘的算法；

(5) 不确定图数据挖掘的应用。

下面，我们分别对这几方面的研究工作进行简要介绍。

1.4.1 不确定图数据模型

不确定图数据模型在不确定图数据挖掘中具有重要作用，它定义了不确定图数据的表示形式及语义，为不确定图数据的表示、存储以及不确定图数据挖掘问题的形式化定义奠定了基础。

本书作者最早在文献 [25] 中扩展了不确定数据库中常用的**可能世界模型** (possible worlds model) [51–55]，提出了不确定图数据的可能世界模型。在该模型中，不确定图的每条边上带有一个 $[0,1]$ 内的实数，称作

存在概率 (existence probability)，表示该边实际存在的概率。该模型假定边上的存在概率之间相互独立。依据不确定图边上的存在概率对边进行独立随机选取，可以得到该不确定图的一种可能的确定图存在形式(即可能世界)，称作**蕴含图** (implicated graph)。图 1.2 给出了一个不确定图，图 1.3 给出了该不确定图的 2^7 个蕴含图中的 6 个。理论证明，一个不确定图代表了其全部蕴含图上的一个概率分布 [27]。后续不确定图数据挖掘和不确定图管理方面的研究 [18, 26–29, 33–44, 46] 基本上均采用了该语义模型。

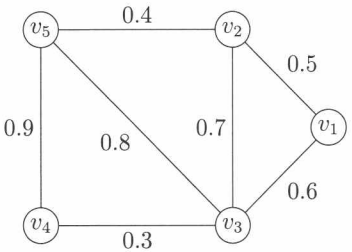


图 1.2　不确定图

本书作者随后在文献 [30–32] 中扩充了不确定图数据的可能世界语义模型，使不确定图的每个顶点上也带有一个 $[0,1]$ 内的实数，表示该顶点实际存在的概率；同时，边上的实数表示在该边的两个端点都已存在的条件下，该边实际存在的条件概率。该模型假定顶点上的存在概率之间相互独立，边上的条件存在概率之间也相互独立。

上述两种可能世界模型均基于不确定性相互独立的假设。为了表示不确定性之间的相互依赖关系，Hua 和 Pei [24] 在不确定图数据的可能世界语义模型的基础上增加了一个条件依赖概率表，用于记录边上的存在概率之间的条件依赖关系。然而，由于边之间的相互依赖关系很难获得，因此后续研究很少使用这种不确定图数据模型。

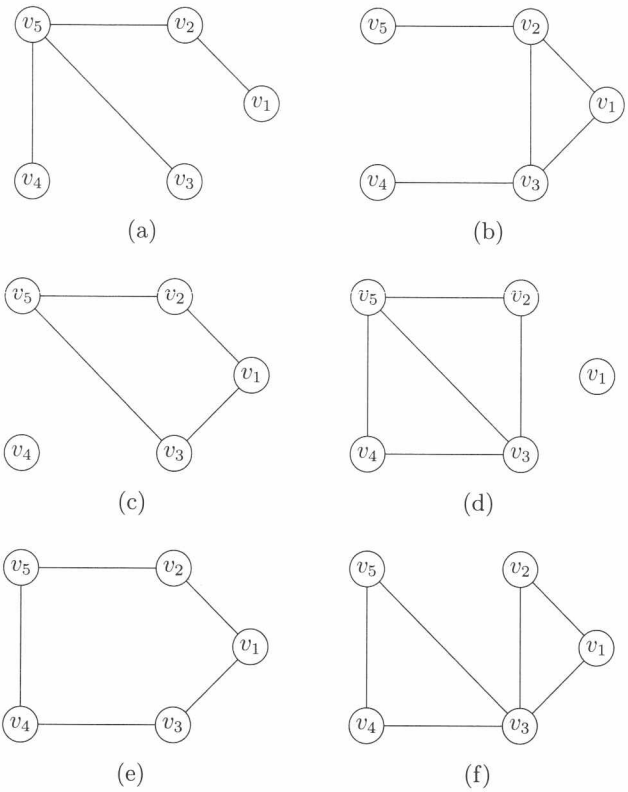


图 1.3　不确定图的部分蕴含图

1.4.2　不确定图数据挖掘问题的语义

由于不确定图数据模型在语义上的特殊性，不确定图数据挖掘问题与确定图数据挖掘问题在语义上存在很大的差异。

确定图数据挖掘问题通常使用一个实函数来度量知识的重要性，并根据该度量函数值来确定输出哪些知识。如图 1.4 所示，一项知识 K 在一个确定图数据 D 中的重要性为一个确定值 $f(K;D) \in \mathbb{R}$。